Z-122E
c.2.

Statistics Canada Statistique Canada

Ottawa

APPROXIMATIONS TO THE DISTRIBUTION OF A SUM OF WEIGHTED CHI-SQUARE VARIABLES

M.A. HIDIROGLOU

JANUARY 1983

TABLE OF CONTENTS

	PAGE
1. Introduction	1
2. Solomon-Stephens Approximation	2
3. Satterthwaite's Approximation	3
4. Uses of the Approximations	4
5. Empirical Results	8
Example 1: (Independence in a 4×2 table)	8
Example 2: (Independence in a 7×2 table)	10
Example 3: (Independence in a 3×3 table)	12
Example 4: (Independence in a 5×3 table)	13
Example 5: (Independence in a 3×3 table)	14
6. Summary	16
7. Acknowledgements	17
Table 1 : Population by Variety of Drugs Taken, by sex in thousands	8
Table 2 : Chi-Square Statistics & Estimated Levels: Nominal Size 0.05	9
Table 3 : Population 5 years & Over by diastolic blood pressure by sex	11
Table 4 : Chi-Square Statistics & Estimated Levels: Nominal Size 0.05	11
Table 5 : Population 15-64 years by Fitness Level & Type of Cigarette Smoker	12
Table 6 : Chi-Square Statistics & Estimated Levels: Nominal Size 0.05	13
Table 7 : Population 15 & over by diastolic blood pressure & type of cigarette smoker	13
Table 8 : Chi-Square Statistics & Estimated Levels: Nominal Size 0.05	14
Table 9 : Population 15 years & over by type of beverage and number of drinks consumed per week	15
Table 10: Chi-Square Statistics & Estimated Levels: Nominal Size 0.05	15
Table 11: Summary Chi-Square Statistics for Examples 1-5	16
Appendix - Table 12: Summary Statistics for Examples 1-5	18
Table 13: Eigenvalues for Examples 1-5	18
Table 14: Frequency Distribution of the Cell Coefficients of Variation for Examples 1-5	19
References	20

APPROXIMATIONS TO THE DISTRIBUTION OF A SUM OF WEIGHTED CHI-SQUARE VARIABLES

1. Introduction

There are many areas in statistical analysis that need probability computations of quadratic forms in normal variables which we denote by X^2. The quadratic form X^2 is of the type

$$X^2 = \sum_{j=1}^{k} \lambda_j (S_j + a_j)^2$$

where S_j are independent and identically distributed standard normal variables (i.e., mean zero and variance one), and where λ_j and a_j are nonnegative constants. This form of linear combination of quadratic forms with $a_j = 0$, $j = 1, \ldots, k$ is one obtained by Rao and Scott (1981) in studying chi-square tests to data obtained from complex surveys. As shown by Rao and Scott (1981), the usual Pearson goodness-of-fit chi-square test,

$$X^2 = n \sum_{i=1}^{k+1} (\hat{p}_i - p_{oi})^2 / p_{oi}$$

where $\hat{p}_i$'s are design-consistent estimator of $\hat{p}_i$'s (the population proportions) under p(s) (the sample design under consideration), with $\Sigma\, \hat{p}_i = 1$, n being the sample size and p_{oi} is the expected proportion under the null hypothesis, may be written for large n as $\sum_{i=1}^{k} \lambda_{oi} S_i^2$. Here S_i's are asymptotically independent N(0,1) random variables and λ_{oi}'s are the eigenvalues of $D_{\underset{\sim}{o}} = n P_{\underset{\sim}{o}}^{-1} \Sigma_{\underset{\sim}{o}}$ ($\lambda_{o1} \geq \lambda_{o2} \geq \ldots \geq \lambda_{o,k} \geq o$), where $\Sigma_{\underset{\sim}{o}}$ denotes the covariance matrix of the $\hat{p}_i$'s under the design and $P_{\underset{\sim}{o}}$ is the covariance matrix of the estimates of p_i's under multinomial sampling, when $p_i = p_{oi}$, $i = 1, \ldots k$.

Significance levels for X^2 are required in order to compare them to a nominal size (α), i.e.

$$SL(X^2) = Pr[X^2 \geq \chi^2_k(\alpha)],$$

where $\chi^2_k(\alpha)$ is the upper α-point of chi-square random variable with k degrees of freedom. These significance levels may be obtained exactly using a procedure given by Imhof (1961). Imhof's procedure is based on numerical inversion of the characteristic function of X^2. This method, though accurate for all practical purposes, is relatively expensive in computer time. Various approximations to $SL(X^2)$ have been proposed: they fall into two types, Gaussian and chi-square. Jensen and Solomon (1972) proposed the Gaussian approximation $Z = (X^2/\theta)^h$, where θ is the mean of X^2, and approximates Z by a normal distribution, where the mean and variance of the approximating distribution depend on the first three cumulants of X^2. Chi-square approximations have been proposed by Satterthwaite (1946) and Solomon and Stephens (1977). Solomon and Stephens' study suggests that their approximation is more accurate than the Jensen-Solomon approximation in both tails of the distribution for X^2. The behaviour of the Solomon-Stephens approximation will be compared to the Satterthwaite approximation in the upper tail.

2. Solomon-Stephens Approximation ($a_j = 0$, $j = 1, \ldots, k$)

The distribution of $X^2 = \Sigma_i \lambda_i S_i^2$ is fitted by $\hat{X}^2 = A Y^r$ where Y has the χ^2_t distribution, and the constants A, t and r are found by matching the first three moments of $\hat{X}^2$ to those of X^2. The first three moments of X^2 when all the a_j's are zero are

$$\mu_1' = \sum_{i=1}^{k} \lambda_i,$$

$$\mu_2' = 2 \sum_{i=1}^{k} \lambda_i^2 + \left(\sum_{i=1}^{k} \lambda_i \right)^2 \qquad (2.1)$$

and

$$\mu_3' = 8 \sum_{i=1}^{k} \lambda_i^3 + 6 \left(\sum_{i=1}^{k} \lambda_i \right) \left(\sum_{j=1}^{k} \lambda_j^2 \right) + \left(\sum_{i=1}^{k} \lambda_i \right)^3$$

while those for $\hat{X}^2$ are

$$\mu_1' = A\,2^r\{\Gamma(r+v)\}/C,$$

$$\mu_2' = A^2 4^r\{\Gamma(2r+v)\}/C, \qquad (2.2)$$

and

$$\mu_3' = A^3 8^r\{\Gamma(3r+v)\}/C,$$

where $v = t/2$ and $C = \Gamma(v)$.

The system of equations (2.1) and (2.2) yields

$$R_2 = \mu_2'/\mu_1'^2 = C\,\Gamma(2r+v)/\{\Gamma(r+v)\}^2 \qquad (2.3)$$

$$R_3 = \mu_3'/\mu_1'^3 = C^2\,\Gamma(3r+v)/\{\Gamma(r+v)\}^3 \qquad (2.4)$$

where R_2 and R_3 are obtained from (2.1). Equations (2.3) and (2.4) can now be solved using Newton-Raphson or secant procedures.

3. Satterthwaite's Approximation

Satterthwaite (1946) fitted X^2 by $\hat{X}^2 = A\,w$ where w has the χ^2_r chi-squared distribution by matching only the first two moments of X^2 to those of $\hat{X}^2$. This approximation provides direct estimates of A and r as

$$A = \left(\sum_{i=1}^{k} \lambda_i^2\right)/\left(\sum_{i=1}^{k} \lambda_i\right)$$

and

$$r = \left(\sum_{i=1}^{k} \lambda_i\right)^2/\left(\sum_{i=1}^{k} \lambda_i^2\right).$$

A and r may be re - expressed as

$$A = \bar{\lambda}\,\{1 + CV^2(\lambda_i)\}$$

and

$$r = k/\{1 + CV^2(\lambda_i)\}$$

where $CV^2(\lambda_i) = \Sigma_i(\lambda_i - \bar{\lambda})^2/k\bar{\lambda}^2$ is the squared coefficient of variation of the λ_i's. For our purposes, the advantage of this approximation over the Solomon-Stephens approximation is that the λ_i's need not be known individually. Rather, only the covariance matrix $\hat{\Sigma}_o$ need be known as will be seen later.

4. Uses of the Approximations

Analysis of survey data is frequently done by summarizing the data in the form of multi-way tables. The summarized data is then analysed using Pearson chi-squared statistics in order to test various hypotheses put forth by social or economic data researchers. The significance levels of such tests are known to be substantially higher than the nominal levels (see Rao and Scott 1981, Holt, Scott and Ewings 1980, Fellegi 1980) if the effect of the sample design is pronounced: that is the combined stratification and cluster effect make the sample design depart significantly from a simple random sample. The effect of design in such cases must be taken into account by either computing a Wald chi-square statistic or modifying the Pearson chi-square statistic. Rao and Scott (1981) have proposed procedures for the latter treatment in the case of goodness-of-fit, homogeneity and independence in two-way tables. The two approximations will be used to study significance levels for the test of independence in particular.

For the test of independence in a two-way table $(I+1) \times (J+1)$, the null hypothesis of interest is given by

$$H_o: \; p_{ij} = p_{i+}\, p_{+j}, \; i=1, 2, \ldots, I+1; \; j=1, 2, \ldots, J+1$$

where p_{ij} denotes the population proportion in the $(i,j)^{th}$ cell, $p_{i+} = \sum_{j=1}^{J+1} p_{ij}$ and $p_{+j} = \sum_{i=1}^{I+1} p_{ij}$. The customary Pearson statistic for testing H_o is

$$X_p^2 = n \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} (\hat{p}_{ij} - \hat{p}_{i+}\, \hat{p}_{+j})^2 / (\hat{p}_{i+}\, \hat{p}_{+j})$$

where $\hat{p}_{ij}$, $\hat{p}_{i+}$ and $\hat{p}_{+j}$ are the respective estimates of p_{ij}, p_{i+} and p_{+j} under the sample design p(s). This test may be written in matrix form as

$$X_p^2 = n\, \hat{\underset{\sim}{h}}^T (\hat{\underset{\sim}{P}}_{I+}^{-1} \otimes \hat{\underset{\sim}{P}}_{+J}^{-1})\, \hat{\underset{\sim}{h}}$$

where $\hat{\underset{\sim}{h}} = (\hat{h}_{11}, \ldots, \hat{h}_{1J}; \ldots; \hat{h}_{I1}, \ldots, \hat{h}_{IJ})^T$ with

$$\hat{h}_{ij} = \hat{p}_{ij} - \hat{p}_{i+}\,\hat{p}_{+j},\quad \hat{\underset{\sim}{P}}_{I+} = \text{diag}\,(\hat{\underset{\sim}{p}}_{I+}) - \hat{\underset{\sim}{p}}_{I+}\,\hat{\underset{\sim}{p}}_{I+}^T,$$

$$\hat{\underset{\sim}{P}}_{+J} = \text{diag}\,(\hat{\underset{\sim}{p}}_{+J}) - \hat{\underset{\sim}{p}}_{+J}\,\hat{\underset{\sim}{p}}_{+J}^T,\quad \hat{\underset{\sim}{p}}_{I+} = (\hat{p}_{1+}, \ldots, \hat{p}_{I+})^T,$$

$$\hat{\underset{\sim}{p}}_{+J} = (\hat{p}_{+1}, \hat{p}_{+2}, \ldots, \hat{p}_{+J})^T,$$

and θ denotes the direct product.

Rao and Scott (1981) have shown that X_p^2 is asymptotically distributed as ($\approx$) a weighted sum, $\sum_{k=1}^{IJ} \delta_{ok}\, w_k$, of independent χ_1^2 random variables w_k under H_o. Here the δ_{ok}'s are the eigenvalues of the "design-effects matrix" $\underset{\sim}{D}_o = n(\underset{\sim}{P}_{I+}^{-1} \theta \underset{\sim}{P}_{+J}^{-1})\,\underset{\sim}{\Gamma}$ where $\underset{\sim}{\Gamma}$ is the asymptotic covariance matrix of $\hat{\underset{\sim}{h}}$. Rao and Scott (1981) proposed a simple modification to X_p^2 requiring only the deffs of the $\hat{h}_{ij}$'s,

$d_{ij}(\hat{\underset{\sim}{h}}) = \text{var}(\hat{h}_{ij})/[\hat{p}_{i+}(1-\hat{p}_{i+})\,\hat{p}_{+j}(1-\hat{p}_{+j})/n]$ i=1, 2, ..., I+1; j=1, 2, ..., J+1, where $\text{var}(\hat{h}_{ij})$ is the estimated variance of $\hat{h}_{ij}$ under H_o. The form of $\text{var}(\hat{h}_{ij})$ has been given by Hidiroglou and Rao (1981) for the case of the estimated proportions being derived as the ratio of unadjusted blown-up counts and as the ratio of adjusted (post-stratified on age-sex) blown-up counts. The modified statistic is

$$X_p^2\,(\hat{\delta}_{\cdot}) = X_p^2 \,/\, \hat{\delta}_{\cdot},$$

where (IJ) $\hat{\delta}_{\cdot} = \sum_{i=1}^{I+1}\sum_{j=1}^{J+1} d_{ij}(\hat{\underset{\sim}{h}})(1-\hat{p}_{i+})(1-\hat{p}_{+j})$.

An improved approximation, but requiring full knowledge of $\hat{\underset{\sim}{\Gamma}}$, is obtained by treating $X_p^2(S) = X_p^2(\hat{\delta}_{\cdot})/(1+\hat{C}^2)$ as χ_ν^2 where

$\nu = IJ/(1+\hat{C}^2)$, $\hat{C}^2 = \sum_{k=1}^{IJ} \hat{\delta}_k^2/(\hat{\delta}_{\cdot}^2\, IJ) - 1$. $\hat{C}^2$ may be obtained as a function of the estimated proportions and the estimated covariance

by using $\sum_{k=1}^{IJ} \hat{\delta}_k^2 = \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} \sum_{m=1}^{I+1} \sum_{\ell=1}^{J+1} \hat{\sigma}_{ij,m\ell}(\hat{\underset{\sim}{h}})/(\hat{p}_{i+}\,\hat{p}_{+j}\,\hat{p}_{m+}\,\hat{p}_{+\ell})$

without evaluating individual $\hat{\delta}_i$'s, where $\hat{\sigma}_{ij,m\ell}(\hat{\underset{\sim}{h}})$ is the estimated covariance between $\hat{h}_{ij}$ and $\hat{h}_{m\ell}$.

The Wald statistic takes the sample design into account and is of the form:

$$X_W^2 = \hat{\underset{\sim}{h}}^T \hat{\underset{\sim}{\Gamma}}^{-1} \hat{\underset{\sim}{h}}$$

where $\hat{\underset{\sim}{\Gamma}}$ is the estimated covariance matrix of $\hat{\underset{\sim}{h}}$ under the particular sample design used.

The asymptotic significance level of X_P^2 under H_o may be obtained using the Satterthwaite approximation as

$$SL\ (X_P^2) = Pr[X_P^2 \geq \chi_{IJ}^2(\alpha)]$$

$$\doteq Pr[\chi_\nu^2 \geq \chi_{IJ}^2(\alpha)/\{\hat{\delta}_\cdot(1+\hat{C}^2)\}]$$

where $\chi_{IJ}^2(\alpha)$ is the upper α - percentage point of a χ^2 random variable with IJ degrees of freedom. Using the Solomon-Stephens approximation, the estimated significance level is

$$SL(X_P^2) = Pr[X_P^2 \geq \chi_{IJ}^2(\alpha)]$$

$$\doteq Pr[\chi_t^2 \geq \{\chi_{IJ}^2(\alpha)/A\}^{1/r}]$$

where X_P^2 has been fitted to the distribution of $A\,Y^r$ with Y distributed as χ_t^2.

Modified chi-square statistics of the form $X_P^2(\hat{b}_\cdot) = X_P^2/\hat{b}_\cdot$ have been proposed in the literature where $\hat{b}_\cdot$ is a suitable linear combination of the estimated cell design effects $\hat{d}_{ij}$. Using Satterthwaite's approximation, the estimated significance level of $X_P^2(\hat{b}_\cdot)$ is

$$SL[X_P^2(\hat{b}_\cdot)] = Pr[X_P^2(\hat{b}_\cdot) \geq \chi_{IJ}^2(\alpha)]$$

$$\doteq Pr[\chi_\nu^2 \geq \hat{b}_\cdot \chi_{IJ}^2(\alpha)/\{\hat{\delta}_\cdot(1+\hat{C}^2)\}] .$$

Using the Solomon-Stephens approximation, the estimated significance level is

$$SL[X_P^2(\hat{b}_\cdot)] = Pr[X_P^2(\hat{b}_\cdot) \geq \chi_{IJ}^2(\alpha)]$$

$$\doteq Pr[\chi_t^2 \geq \{\hat{b}_\cdot \chi_{IJ}^2(\alpha)/A\}^{1/r}] .$$

Two particular cases of $X_P^2(\hat{b}_\cdot)$ are:

$$X_P^2(\hat{d}_\cdot) = X_P^2/\hat{d}_\cdot$$

$$X_P^2(\hat{\lambda}_\cdot) = X_P^2/\hat{\lambda}_\cdot$$

where

$$\hat{d}_\cdot = \sum_{i=1}^{I+1}\sum_{j=1}^{J+1} \hat{d}_{oij}/[(I+1)(J+1)]$$

and

$$\hat{\lambda}_\cdot = \sum_{i=1}^{I+1}\sum_{j=1}^{J+1} (1 - \hat{p}_{i+}\hat{p}_{+j})\,\hat{d}_{oij}/[(I+1)(J+1) - 1]$$

with $\hat{d}_{oij}$ being the estimated design effect of $\hat{p}_{ij}$ under H_o, that is: $\hat{d}_{oij} = var(\hat{p}_{ij})/[\hat{p}_{i+}\hat{p}_{+j}(1 - \hat{p}_{i+}\hat{p}_{+j})/n]$.

A modification based on Satterthwaite's approximation is given by

$$X_P^2(S,\alpha) = X_P^2\left[\frac{\chi_{IJ}^2(\alpha)}{\hat{\delta}_\cdot(1+\hat{C}^2)\,\chi_\nu^2(\alpha)}\right]$$

where $SL[X_P^2(S,\alpha)] = P[X_P^2(S,\alpha) \geq \chi_{IJ}^2(\alpha)]$

$$\doteq \alpha .$$

Using the Solomon-Stephens approximation,

$$X_p^2(ST) = (X_p^2/A)^{1/r}$$

and

$$X_p^2(ST,\alpha) = X_p^2(ST)[\chi_{IJ}^2(\alpha)/\chi_t^2(\alpha)]$$

is such that $SL[X_p^2(ST,\alpha)] \doteq \alpha$.

5. Empirical Results

We now investigate the large-sample performance of X_p^2 for the test of independence and its given modifications, using some data from the Canada Health Survey. This survey is a stratified multi-stage cluster sample design (see The Health of Canadians pp. 229-233). All the computations have been done for the case of proportions of age-sex adjusted totals: further details of the estimation procedures used can be found in Hidiroglou and Rao (1981).

- Example 1: (Independence in a 4×2 table)

Consider the estimated counts cross-classified by drug use (four categories: 0, 1, 2, 3+ drug classes in a 2-day period) and sex (male, female). Here, n = 31,688, I+1 = 4 and J+1 = 2 (IJ = 3) Drug use includes information on the use of medicines, pills or ointments within the past two days. The estimated counts (in thousands) of the population reporting in each of the drug by sex categories are given in the table provided below.

Table 1: Population by Variety of Drugs Taken, by sex (in thousands)

Sex	Drugs			
	0	1	2	3+
Male	6759	3081	1100	476
	(0.011)a	(0.026)	(0.027)	(0.045)
	(1.56)b	(3.37)	(1.15)	(1.38)
Female	5243	3659	1669	1035
	(0.020)	(0.023)	(0.034)	(0.036)
	(3.59)	(3.13)	(2.85)	(1.96)

Note: "a" stands for coefficient of variation of the cell
"b" stands for the cell DEFF

FOOTNOTE: A star above the numbers indicates that these numbers were suppressed in the Health of Canadians (1981) report.

Table 2: Chi-Square Statistics and Estimated Levels: Nominal Size 0.05

Statistic	Estimated Level Using	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 774$	0.226	0.216
$X_p^2(\hat{\delta}_\cdot) = 437$	0.062	0.060
$X_p^2(\hat{\lambda}_\cdot) = 331$	0.022	0.024
$X_p^2(\hat{d}_\cdot) = 327$	0.023	0.023
$X_p^2(S,0.05) = 408$	0.050	0.049
$X_p^2(ST) = 189$	-	-
$X_p^2(ST,0.05) = 138$	-	0.050
$X_W^2 = 538$	-	-

Note: Dashes indicate that no computation was done.

Table 2 gives the estimated significance levels using the Satterwaite approximation and the Solomon-Stephens approximation. For this example the eigenvalues were 0.9982, 1.3728 and 2.9386 with an estimated coefficient of variation equal to 0.47. The constants in the Solomon-Stephens approximation are: $r = 1.38$, $A = 0.558$ and $t = 4.766$. In this example, the Pearson statistic is quite high (774). The Wald statistic which is distributed approximately as a χ_3^2 random variable under H_o is quite high (538) relative to the modified Pearson statistics. The $\hat{\delta}_\cdot$ modification provides an estimated level fairly close to the required nominal level (0.05). Note, however, that both the $\hat{\lambda}_\cdot$ and $\hat{d}_\cdot$ modifications are quite conservative. The estimated significance levels using either the Solomon-Stephens or Satterthwaite approximations are quite close to each other.

Note on the other hand that the chi-square statistic obtained using the Solomon-Stephens approximation is quite a bit lower

(189) than all the other provided statistics. In fact, $X_p^2(ST) = 189$ is smaller that $X_p^2/(\text{max eigenvalue}) = 263.4$. This is something not to be expected since $X_p^2 = \sum_{k=1}^{IJ} \delta_{ok} W_k \leq \delta_{max} (\sum_{k=1}^{IJ} W_k)$ where W_k's are independent $\chi^2_{(1)}$ random variables and $\delta_{max} = \max \{\delta_{ok}: k = 1, 2, \ldots, IJ\}$. Now, $(X_p^2/\delta_{max}) \leq \sum_{k=1}^{IJ} W_k \approx \chi^2_{(IJ)}$; hence, $P[X_p^2/(\delta_{max}) \geq \chi^2_{IJ}(\alpha)] \leq \alpha$. The explanation for this phenomenon is that the Solomon-Stephens approximation is not accurate in the extreme tails and the value $X_p^2 = 774$ is quite far out in the tail of the distribution with a minutely small α.

An approximation based on three moments (the first, second and third) is adequate over most of the range but not necessarily in the extreme tails. Box's (1954) investigation, however, has shown that the Satterthwaite approximation is fairly accurate in the tail of interest. In our example, $X_p^2(S,.05)$ would have been equal to $X_p^2(ST,.05)$ if X_p^2 had been equal to 15.1. This is obtained by equating $X_p^2(S,0.05)$ to $X_p^2(ST,.05)$ and solving for X_p^2:

$$\text{i.e. solving } X_p^2 = \exp\{\frac{r}{r-1}\,[\ell n\, d - \frac{1}{r}\,\ell n\, A - \ell n\, g]\}$$

where $d = \chi^2_{IJ}(\alpha)/\chi^2_t(\alpha)$ and $g = \dfrac{\chi^2_{IJ}(\alpha)}{\hat{\delta}_{\cdot}(1+\hat{c}^2)\,\chi^2_{\nu}(\alpha)}$ with $\alpha = 0.05$, $\nu = 2.45$, $IJ = 3$, $\hat{\delta}_{\cdot} = 1.77$ and $\hat{c} = 0.47$.

- Example 2: (Independence in a 7×2 table)

Consider the estimated counts cross-classified by diastolic blood pressure (seven categories: less than 55, 55-64, 65-74, 75-84, 85-94, 95-104, 105+ in units of mmHg) and sex (male, female). Here n = 5760, I+1 = 7 and J+1 = 2. Blood pressure was measured on respondents five years of age and older. The estimated counts

(in thousands) for the above cross-classification are provided in the table provided below.

Table 3: Population 5 years and over by diastolic blood pressure by sex

Sex	Blood Pressure						
	<55	55-64	65-74	75-84	85-94	95-104	105+
Male	390	1307	2589	3524	1905	544	184
	(0.124)a	(0.072)	(0.065)	(0.046)	(0.070)	(0.154)	(0.203)
	(1.69)b	(1.98)	(3.35)	(2.31)	(2.87)	(3.67)	(2.11)
Female	484	1900	3490	2981	1270	478	44*
	(0.134)	(0.060)	(0.039)	(0.063)	(0.080)	(0.173)	(0.231)
	(2.56)	(1.94)	(1.76)	(3.70)	(2.33)	(4.03)	(0.64)

Note: "a" stands for coefficient of variation of the cell
"b" stands for the cell DEFF

Table 4: Chi-Square Statistics and Estimated Levels: Nominal Size 0.05

Statistic	Estimated Level Using	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 141$	0.307	0.290
$X_p^2(\hat{\delta}_\cdot) = 81$	0.081	0.077
$X_p^2(\hat{\lambda}_\cdot) = 56$	0.019	0.021
$X_p^2(\hat{d}_\cdot) = 56$	0.018	0.021
$X_p^2(S,0.05) = 70$	0.050	0.049
$X_p^2(ST) = 59$	-	-
$X_p^2(ST,0.05) = 36$	-	0.050
$X_W^2 = 107$	-	-

Note: Dashes indicate that no computation was done

Table 4 provides the estimated significance levels using the Satterthwaite approximation and Solomon-Stephens approximation for the usual Pearson statistic X_p^2 and its modifications. Summary conclusions for the above table will be provided after all examples have been produced.

- Example 3: (Independence in a 3×3 table)

Physical fitness was measured from pulse readings obtained after three minutes of a stepping exercise. Respondents were classified into three categories: "Recommended Level" (pulse below specified rate after six minutes), "Minimum Acceptable" (pulse rate below three-minute criterion but above six-minute criterion) and "Unacceptable" (pulse rate above criterion at three minutes). Physical fitness was classified by smoking level. The analysis was done for the domain where the fitness level was considered to be above a required minimum. Three levels of smoking were defined: current smoker, occasional smoker, non-smoker. Here n = 1731, I+1 = 3 and J+1 = 3. The estimated counts (in thousands) are provided below.

Table 5: Population 15-64 years by Fitness Level and Type of Cigarette Smoker

Type of Smoker	Fitness Level		
	Recommended	Minimum	Below
Current	2083 (0.083)a (4.28)b	1415 (0.091) (3.26)	138* (0.230) (2.76)
Occasional	218* (0.246) (4.06)	91* (0.182) (0.86)	8* (0.801) (1.68)
Never	1924 (0.081) (1.86)	940 (0.071) (1.26)	32* (0.314) (0.77)

NOTE: "a" stands for coefficient of variation of the cell
"b" stands for the cell DEFF

Table 6: Chi-Square Statistics and Estimated Levels: Nominal Size 0.05

Statistic	Estimated Level Using	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 24$	0.138	0.130
$X_p^2(\hat{\delta}_{\cdot}) = 18$	0.071	0.068
$X_p^2(\hat{\lambda}_{\cdot}) = 11$	0.008	0.009
$X_p^2(\hat{d}_{\cdot}) = 10$	0.006	0.007
$X_p^2(S,0.05) = 16$	0.050	0.049
$X_p^2(ST) = 20$	-	-
$X_p^2(ST,0.05) = 14$	-	0.050
$X_W^2 = 24$	-	-

NOTE: Dashes indicate that no computation was done.

- Example 4: (Independence in a 5×3 table)

The table to be studied is cross-classified by diastolic blood pressure (five categories: 55-64, 65-74, 75-84, 85-94, 95-104 in units of mmHg) and type of smoking habit (three categories: current smoker, past smoker, never smoke). Here $n = 4007$, $I+1 = 5$ and $J+1 = 3$. The estimated counts (in thousands) for the above cross-classification are provided in the table provided below.

Table 7: Population 15 and over by diastolic blood pressure and type of cigarette smoker

Type of Smoker	Blood Pressure				
	55-64	65-74	75-84	85-94	95-104
Current	824 (0.096)a (2.19)b	1876 (0.055) (1.92)	2087 (0.054) (2.03)	943 (0.091) (2.38)	364 (0.195) (3.72)
Past	384 (0.100) (1.12)	1073 (0.108) (3.99)	1574 (0.116) (5.53)	834 (0.098) (1.98)	256 (0.177) (2.07)
Never	475 (0.136) (2.41)	1417 (0.066) (1.78)	1817 (0.072) (2.36)	885 (0.064) (1.11)	293 (0.161) (2.10)

Table 8: Chi-Square Statistics and Estimated Levels: Nominal Size 0.05

Statistic	Estimated Level Using	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 31$	0.308	0.289
$X_p^2(\hat{\delta}_{\cdot}) = 19$	0.107	0.099
$X_p^2(\hat{\lambda}_{\cdot}) = 13$	0.023	0.024
$X_p^2(\hat{d}_{\cdot}) = 13$	0.022	0.024
$X_p^2(S,0.05) = 15$	0.050	0.049
$X_p^2(ST) = 24$	-	-
$X_p^2(ST,0.05) = 15$	-	0.050
$X_W^2 = 36$	-	

- Example 5: (Independence in a 3×3 table)

The final example provides a cross-classification between type of beverage (beer, wine, liquor) and number of drinks consumed per week (1-5, 6-13, 14+ drinks). Here n = 6966, I+1 = 3 and J+1 = 3. The estimated counts (in thousands) are provided below.

Table 9: Population 15 years (current drinkers only) and over by type of beverage (usually consumed) and number of drinkgs consumed per week

Number of drinks	Beverage		
	Beer	Wine	Liquor
1-5	223 (0.082)a (1.28)b	1124 (0.038) (1.15)	4585 (0.026) (0.62)
6-13	327 (0.060) (1.00)	806 (0.049) (1.04)	777 (0.048) (1.23)
14+	13* (0.258) (0.61)	1162 (0.045) (1.94)	294 (0.063) (1.01)

Table 10: Chi-Square Statistics and Estimated Levels: Nominal Size 0.05

Statistic	Estimated Level Using	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 1974$	0.255	0.258
$X_p^2(\hat{\delta}_\cdot) = 1102$	0.072	0.072
$X_p^2(\hat{\lambda}_\cdot) = 1763$	0.210	0.210
$X_p^2(\hat{d}_\cdot) = 1765$	0.218	0.218
$X_p^2(S,0.05) = 970$	0.050	0.049
$X_p^2(ST) = 336$	-	-
$X_p^2(ST,0.05) = 282$	-	0.050
$X_W^2 = 660$	-	-

6. Summary

The values for the X_P^2 chi-square statistic and selected modified versions for the five data sets are summarized in table 11. Several points may be noted from this table. For all the examples chosen, all the chi-square statistics are significant to at least the 5% level. The Wald statistic appears to be consistently higher than the given modified versions of the Pearson chi-square statistics and smaller than the Pearson chi-square statistic (see examples 1-4).

Table 11: Summary Chi-square Statistics for Examples 1-5

Example	Chi-square Statistic				
	X_P^2	$X_P^2(\hat{\delta}_\cdot)$	$X_P^2(S,0.05)$	$X_P^2(ST,0.05)$	X_W^2
1. DRUG*SEX	774***	437***	408***	138***	538***
2. PRESS*SEX	141***	81***	70***	36***	107***
3. FIT*SMOK	24***	18***	16***	14***	24***
4. PRESS*SMOK	31***	19**	15**	15**	36***
5. BEV*DRINK	1974***	1102***	970***	282***	660***

** - 0.05 level of significance
*** - 0.01 level of significance

The modified Solomon-Stephens chi-square statistics, $X_P^2(ST,0.05)$ is consistently lower than the other chi-square statistics. In fact, when the Pearson chi-square value is very high (see examples 1, 2 or 5), the modified Solomon-Stephens chi-square statistics is much lower than the other statistics. When the Pearson chi-square statistic is not high (as in examples 3 or 4), the modified Solomon-Stephens chi-square statistic is much closer to the modified Satterthwaite statistic $X_P^2(S,0.05)$. Note also that the estimated nominal levels obtained using either the Satterthwaite or Solomon-Stephens approximations are quite close to each other.

$X_P^2(S,0.05)$, $X_P^2(ST,0.05)$ or X_W^2 require the computation of the covariance matrix of the cell proportion estimates. For small contingency tables, this need not be a problem. For larger tables, as the

number of variances and covariances increases with the square of the number of cells, which itself increases with the product of the number of categories in each classification variable, this matrix quickly gets out of hand. Furthermore, X^2_W requires the inversion of the estimated covariance matrix which may itself be instable. $X^2_p(ST,0.05)$ moreover requires the computation of eigenvalues of the design effect matrix $\underset{\sim}{D}_o$. Note that $X^2_p(\hat{\delta}_.)$ is close to $X^2_p(S,0.05)$ in terms of value and estimated nominal level and requires far fewer computations to be obtained.

For all the examples provided in this report, inferences based on all approaches yield the same result: the null hypothesis is rejected. Acceptance or rejection of the null hypothesis using $X^2_p(S,0.05)$ depends on two factors, $\hat{\delta}_.$ and X^2_p. There are probably instances where the modified statistics do not yield the same inferences (for borderline cases). It would be interesting to compare the behaviour of all proposed chi-square statistics using a Monte-Carlo study. Such a study would provide information on the actual significance levels as opposed to the nominal ones for all the proposed statistics under various conditions. Some of these conditions are as follows: How suceptible are the chi-square statistics to variability amongst the individual cell designs effects in terms of inference? What is the effect of the magnitude of large or small X^2_p on the inference? Small-sample power considerations would also be important.

7. Acknowledgements

The author is most thankful to Philip Giles for his assistance in preparing summary tapes. Thanks are also due to Nanjamma Chinnappa and Geoffrey Hole for their support. The author is most indebted to Professor J.N.K. Rao for his suggestions and comments.

APPENDIX

Table 12: Summary Statistics for Examples 1-5

Statistic	Example				
	1 DRUG*SEX	2 PRESS*SEX	3 FIT*SMOK	4 PRESS*SMOK	5 BEV*DRINK
A	0.560	0.075	0.233	0.062	0.693
r	1.38	1.85	1.54	1.94	1.37
t	4.77	13.70	6.98	14.63	5.17
$\hat{d}_.$	2.37	2.50	2.31	2.45	1.09
$\hat{\lambda}_.$	2.34	2.49	2.18	2.43	1.12
$\hat{\delta}_.$	1.77	1.74	1.28	1.61	1.79
$cv(\hat{\lambda}_i)$	0.47	0.75	0.63	1.08	0.69
ν	2.45	3.84	2.86	3.67	2.71
IJ	3.00	6.00	4.00	8.00	4.00
n	31688	5760	1731	4007	6966

Table 13: Eigenvalues for Examples 1-5

Example	Eigenvalues ($\hat{\delta}_i$)							
	1	2	3	4	5	6	7	8
1. DRUG*SEX	1.00	1.37	2.94					
2. PRESS*SEX	0.51	0.73	0.82	1.89	2.19	4.32		
3. FIT*SMOK	0.45	0.75	1.36	2.56				
4. PRESS*SMOK	0.20	0.42	0.47	0.69	0.92	1.20	3.65	5.36
5. BEV*DRINK	0.26	1.15	2.17	3.58				

Table 14: Frequency Distribution of the Cell Coefficients of Variation for Examples 1-5

Example	Coefficient of Variation Ranges				Mean (cv)	Sample Size n
	[0-5]	(5-10]	(10-20]	(20+		
1. DRUG*SEX	100%	-	-	-	3%	31688
2. PRESS*SEX	14%	43%	29%	14%	11%	5760
3. FIT*SMOK	-	57%	11%	44%	23%	1731
4. PRESS*SMOK	-	60%	40%	-	10%	4007
5. BEV*DRINK	56%	33%	-	11%	7%	6966

References

BOX, G.E.P., Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification, Ann.Math.Stat., Vol. 25 (1954), pp. 290-302.

FELLEGI, I.P. (1980), Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples, Journal of the American Statistical Association, 75, 261-268.

HIDIROGLOU, M.A. and RAO, J.N.K. (1981), Chisquare Tests for the Analysis of Categorical Data from the Canada Health Survey, presented at the 43rd ISI Conference in Buenos Aires.

HOLT, D., SCOTT, A.J., and EWINGS, P.O. (1980), Chi-Squared Tests with Survey Data, Journal of the Royal Statistical Society, Sec. A, 143, 302-320.

IMHOF, J.P., (1961), Computing the Distribution of Quadratic Forms in Normal Variables, Biometrika, 417-426.

JENSEN, Donald R., and SOLOMON, Herbert (1972), A Gaussian Approximation to the Distribution of a Definite Quadratic Form, Journal of the American Statistical Association, 67, 898-902.

RAO, J.N.K. and SCOTT, A.J. (1981), The Analysis of Categorical Data from Complex Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, Journal of the American Statistical Association, 76, 221-230.

SATTERTHWAITE, F.E. (1946), An Approximate Distribution of Estimates of Variance Components, Biometrics, 2, 110-114.

SOLOMON, H. and STEPHENS, M.A. (1977), Distribution of a Sum of Weighted Chi-Square Variables, Journal of the American Statistical Association, 72, 881-885.

The Health of Canadians, (1981), Report of the Canada Health Survey, Health and Welfare, and Statistics Canada publication, Catalogue 82-538E.

Z-122F
c.2

Statistics Canada
Statistique Canada

Ottawa

APPROXIMATIONS DE LA DISTRIBUTION D'UNE SOMME PONDÉRÉE DE VARIABLES KHI-CARRÉ

M.A. Hidiroglou
octobre 1982

TABLE DES MATIÈRES

PAGE

1. Introduction 1
2. Approximation de Solomon-Stephens 3
3. Approximation de Satterthwaite 4
4. Utilisations des approximations 5
5. Résultats empiriques 12
   Exemple 1 : (Test d'indépendance dans un tableau 4 x 2) 12
   Exemple 2 : (Test d'indépendance dans un tableau 7 x 2) 15
   Exemple 3 : (Test d'indépendance dans un tableau 3 x 3) 17
   Exemple 4 : (Test d'indépendance dans un tableau 5 x 3) 19
   Exemple 5 : (Test d'indépendance dans un tableau 3 x 3) 20
6. Résumé 22
7. Remerciements 24

Tableau 1 : Population totale selon la variété de médicaments consommés et le sexe (en milliers) 13
Tableau 2 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05 13
Tableau 3 : Population de 5 ans et plus selon la tension artérielle diastolique et le sexe 16
Tableau 4 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05 16
Tableau 5 : Population de 15 à 64 ans selon le niveau de condition physique et le genre de fumeur de cigarettes ...18
Tableau 6 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05 18
Tableau 7 : Population de 15 ans et plus selon la tension artérielle diastolique et le genre de fumeur de cigarettes 19
Tableau 8 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05 20
Tableau 9 : Population de 15 ans et plus selon le genre de boisson consommée et le nombre de consommations par semaine 21
Tableau 10 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05 21
Tableau 11 : Résumé des variables khi-carré des exemples 1 à 5 22

ANNEXE - Tableau 12 : Renseignements sommaires sur les exemples 1 à 5 .25
Tableau 13 : Valeurs propres des exemples 1 à 5 25
Tableau 14 : Distribution de fréquence des coefficients de variation dans chaque case, exemples 1 à 5 ...26

Bibliographie 27

APPROXIMATIONS DE LA DISTRIBUTION D'UNE SOMME PONDÉRÉE DE VARIABLES KHI-CARRÉ

1. Introduction

Il existe un grand nombre d'analyses statistiques où il faut calculer des probabilités pour des formes quadratiques qui contiennent des variables normales que nous appelons X^2. La forme quadratique X^2 est du type

$$X^2 = \sum_{j=1}^{k} \lambda_j (S_j + a_j)^2$$

où les S_j sont des variables normales réduites (à moyenne nulle et à variance égale à l'unité) qui sont indépendantes et dont la distribution est identique, tandis que λ_j et a_j sont des constantes non négatives. Rao et Scott (1981) ont examiné ce genre de combinaison linéaire de formes quadratiques lorsque $a_j = 0$, $j = 1, \ldots, k$, dans une étude de tests khi-carré appliqués aux données d'enquêtes complexes. Ces auteurs ont démontré que le test khi-carré habituel de Pearson pour vérifier la validité de l'ajustement,

$$X^2 = n \sum_{i=1}^{k+1} (\hat{p}_i - p_{oi})^2 / p_{oi}$$

où les $\hat{p}_i$ sont les estimations des $\hat{p}_i$ (proportions de la population) adaptées au plan d'échantillonnage $p(s)$, $\Sigma \hat{p}_i = 1$, n est la taille de l'échantillon et p_{oi} est la proportion prévue en vertu de l'hypothèse nulle, peut s'écrire sous la forme $\sum_{i=1}^{k} \lambda_{oi} S_j^2$ quand la valeur de n est

élevée. Dans ce cas, les S_i sont des variables aléatoires asymptotiquement indépendantes dont la distribution est de type N(0,1) et les λ_{oi} sont les valeurs propres de $\underset{\sim}{D}_o = n \underset{\sim}{P}_o^{-1} \underset{\sim}{\Sigma}_o$ ($\lambda_{oi} \geq \lambda_{o2} \geq \ldots \geq \lambda_{o,k} \geq o$), où $\underset{\sim}{\Sigma}_o$ représente la matrice des covariances des $\hat{p}_i$ du plan d'échantillonnage et $\underset{\sim}{P}_o$ est la matrice des covariances des estimations de p_i calculées à partir d'un échantillon multinomial, quand $p_i = p_{oi}$, $i = 1, \ldots, k$.

On doit ensuite connaître les seuils de signification (SS) pour pouvoir comparer une valeur de X^2 à une valeur critique de khi-carré $\chi^2_k(\alpha)$, c'est-à-dire :

$$SS(X^2) = Pr[X^2 \geq \chi^2_k(\alpha)],$$

où $\chi^2_k(\alpha)$ est la valeur d'une variable aléatoire khi-carré à k degrés de liberté au début d'une région critique qui couvre une proportion α de la surface à la droite de la distribution. Une méthode proposée par Imhof (1961) permet de déterminer exactement ces seuils de signification par l'inversion numérique de la fonction caractéristique de X^2. Bien que cette méthode de calcul soit exacte dans toutes les applications pratiques, elle coûte relativement cher à cause du temps machine qu'elle requiert. Diverses approximations de $SS(X^2)$ ont été proposées; elles se classent en deux catégories, les approximations gaussiennes et les approximations khi-carré. Jensen et Solomon (1972) ont préconisé l'approximation gaussienne $Z = (X^2/\theta)^h$, où θ est la moyenne de X^2, et produit une approximation de Z à l'aide d'une distribution normale où la moyenne et la variance dépendent des trois premiers cumulants de X^2. Des

approximations khi-carré ont été décrites par Satterthwaite (1946) et Solomon et Stephens (1977). L'étude de Solomon et Stephens permet de croire que l'approximation de ces deux auteurs est plus exacte que celle de Jensen-Solomon aux deux extrémités de la distribution de X^2. Plus bas, on compare le comportement de l'approximation de Solomon-Stephens et celui de l'approximation de Satterthwaite à l'extrémité droite de la distribution de khi-carré.

2. Approximation de Solomon-Stephens ($a_j = 0$, $j = 1, \ldots, k$)

La distribution de $X^2 = \Sigma_i \lambda_i S_i^2$ est ajustée par $\hat{X}^2 = A Y^r$, où Y suit une loi de χ^2_t et les constantes A, t et r associées à $\hat{X}^2$ correspondent au trois premiers moments de X^2. Quand tous les a_j sont nuls, les trois premiers moments de X^2 sont :

$$\mu_1' = \sum_{i=1}^{k} \lambda_i,$$

$$\mu_2' = 2 \sum_{i=1}^{k} \lambda_i^2 + \left(\sum_{i=1}^{k} \lambda_i \right)^2 \qquad (2\ 1)$$

et

$$\mu_3' = 8 \sum_{i=1}^{k} \lambda_i^3 + 6\left(\sum_{i=1}^{k} \lambda_i \right)\left(\sum_{j=1}^{k} \lambda_j^2 \right) + \left(\sum_{i=1}^{k} \lambda_i \right)^3$$

alors que ceux de X^2 sont :

$$\mu_1' = A\, 2^r\{\Gamma(r+v)\}/C,$$
$$\mu_2' = A^2 4^r\{\Gamma(2r+v)\}/C, \qquad (2.2)$$

et

$$\mu_3' = A^3 8^r\{\Gamma(3r+v)\}/C,$$

où $v = t/2$ and $C = \Gamma(v)$.

Les systèmes d'équations (2.1) et (2.2) conduisent au résultats suivants :

$$R_2 = \mu_2'/\mu_1'^2 = C\,\Gamma(2r+v)/\{\Gamma(r+v)\}^2 \qquad (2.3)$$
$$R_3 = \mu_3'/\mu_1'^3 = C^2\,\Gamma(3r+v)/\{\Gamma(r+v)\}^3 \qquad (2.4)$$

où R_2 et R_3 sont calculés à partir de (2.1). On peut maintenant résoudre les équations (2.3) et (2.4) au moyen de la technique de Newton-Raphson ou des méthodes de sécantes.

3. Approximation de Satterthwaite

Satterthwaite (1946) a ajusté X^2 à partir de $\hat{X}^2 = A\,w$, où w suit une loi de χ^2_r, en appariant seulement les deux premiers moments de X^2 à ceux de $\hat{X}^2$. Cette approximation fournit des estimations directes de A et r :

$$A = \left(\sum_{i=1}^{k} \lambda_i^2\right) / \left(\sum_{i=1}^{k} \lambda_i\right)$$

et

$$r = \left(\sum_{i=1}^{k} \lambda_i\right)^2 / \left(\sum_{i=1}^{k} \lambda_i^2\right).$$

A et r peuvent se re-écrire:

$$A = \bar{\lambda} \{1 + CV^2(\lambda_i)\}$$

et

$$r = k/\{1 + CV^2(\lambda_i)\}$$

où $CV^2(\lambda_i) = \Sigma_i(\lambda_i - \bar{\lambda})^2/k\bar{\lambda}^2$ est le carré du coefficient de variation des λ_i. Pour nos besoins, l'avantage de cette approximation par rapport à celle de Solomon-Stephens est qu'il n'est pas nécessaire de connaître la valeur précise de chaque λ_i. Il faut seulement connaître la matrice des covariances $\hat{\underset{\sim}{\Sigma}}_o$, comme on le verra plus loin.

4. Utilisations des approximations

L'analyse de données d'enquêtes se fait souvent à l'aide de tableaux à plusieurs dimensions qui résument ces données. Ces tableaux sont ensuite analysés au moyen de variables khi-carré de Pearson afin de vérifier diverses hypothèses élaborées par des spécialistes qui étudient les données sociales ou économiques. Il est reconnu que les seuils de

signification de ce genre de test sont beaucoup plus grands que la région critique choisie (voir Rao et Scott 1981, Holt, Scott et Ewings 1980, Fellegi 1980) si l'effet du plan de sondage est prononcé. Autrement dit, l'effet combiné de la stratification et de la division en grappes conduit à un plan d'échantillonnage très différent d'un plan fondé sur l'échantillonnage aléatoire simple. L'effet du plan doit alors être pris en compte soit par le calcul d'une variable khi-carré de Wald ou une modification de la variable khi-carré de Pearson. Rao et Scott (1981) ont proposé une façon d'appliquer cette dernière solution aux tests de validité de l'ajustement, d'homogénéité et d'indépendance dans les tableaux à deux dimensions. On examine maintenant les seuils de signification des approximations de Solomon-Stephens et de Satterthwaite, en particulier en ce qui a trait au test d'indépendance.

Pour un test d'indépendance dans un tableau à deux dimensions qui contient (I+1) rangées et (J+1) colonnes, l'hypothèse nulle peut se formuler de la façon suivante :

$$H_o : p_{ij} = p_{i+}\, p_{+j},\ i=1, 2, \ldots, I+1;\ j=1, 2, \ldots, J+1$$

où p_{ij} représente la proportion de la population qui se trouve dans la $(i,j)^{\text{ieme}}$ case, $p_{i+} = \sum_{j=1}^{J+1} p_{ij}$ et $p_{+j} = \sum_{i=1}^{I+1} p_{ij}$. La variable khi-carré de Pearson normalement utilisée pour un test de H_o est

$$X_P^2 = n \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} (\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2/(\hat{p}_{i+}\hat{p}_{+j})$$

où $\hat{p}_{ij}$, $\hat{p}_{i+}$ et $\hat{p}_{+j}$ sont les estimations respectives de p_{ij}, p_{i+} et p_{+j} en vertu du plan d'échantillonnage p(s). En notation matricielle, cette variable peut s'écrire :

$$X^2_p = n \, \hat{\underset{\sim}{h}}^T (\hat{\underset{\sim}{P}}^{-1}_{I+} \otimes \hat{\underset{\sim}{P}}^{-1}_{+J}) \, \hat{\underset{\sim}{h}}$$

où $\hat{\underset{\sim}{h}} = (\hat{h}_{11}, \ldots, \hat{h}_{1J}; \ldots; \hat{h}_{I1}, \ldots, \hat{h}_{IJ})^T$ avec

$$\hat{h}_{ij} = \hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j}, \quad \hat{\underset{\sim}{P}}_{I+} = \text{diag}\,(\hat{\underset{\sim}{p}}_{I+}) - \hat{\underset{\sim}{p}}_{I+} \hat{\underset{\sim}{p}}^T_{I+},$$

$$\hat{\underset{\sim}{P}}_{+J} = \text{diag}\,(\hat{\underset{\sim}{p}}_{+J}) - \hat{\underset{\sim}{p}}_{+J} \hat{\underset{\sim}{p}}^T_{+J}, \quad \hat{\underset{\sim}{p}}_{I+} = (\hat{p}_{1+}, \ldots, \hat{p}_{I+})^T,$$

$$\hat{\underset{\sim}{p}}_{+J} = (\hat{p}_{+1}, \hat{p}_{+2}, \ldots, \hat{p}_{+J})^T,$$

et $\otimes$ représente le produit direct.

Rao et Scott (1981) ont démontré que X^2_p est distribué asymptotiquement comme (*) une somme pondérée $\sum_{k=1}^{IJ} \delta_{ok} w_k$, de variables indépendantes w_k de type χ^2_1 selon H_o. Les δ_{ok} sont les valeurs propres de la "matrice des effets du plan" $\underset{\sim}{D}_o = n(\underset{\sim}{P}^{-1}_{I+} \otimes \underset{\sim}{P}^{-1}_{+J}) \, \underset{\sim}{\Gamma}$ où $\underset{\sim}{\Gamma}$ est la matrice des covariances asymptotiques de $\hat{\underset{\sim}{h}}$. Rao et Scott (1981) ont proposé une modification simple de X^2_p qui repose seulement sur l'estimation de l'effet du plan

d'échantillonnage dans les cases

$d_{ij}(\hat{\underset{\sim}{h}}) = var(\hat{h}_{ij})/[\hat{p}_{i+}(1 - \hat{p}_{i+})\ \hat{p}_{+j}(1 - \hat{p}_{+j})/n]$ i=1, 2, ..., I+1; j=1, 2, ..., J+1, où $var(\hat{h}_{ij})$ est l'estimation de la variance de $\hat{h}_{ij}$ en vertu de H_o. La forme de $var(\hat{h}_{ij})$ a été définie par Hidiroglou et Rao (1981) pour le cas où des proportions sont estimées sous la forme d'un rapport de totaux non corrigés de la population et d'un rapport de totaux corrigés de la population (pour tenir compte de la stratification a posteriori en fonction de l'âge et du sexe). La variable modifiée est :

$$X_p^2(\hat{\delta}_.) = X_p^2 / \hat{\delta}_.,$$

où (IJ) $\hat{\delta}_. = \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} d_{ij}(\hat{\underset{\sim}{h}})(1 - \hat{p}_{i+})(1 - \hat{p}_{+j})$.

On peut obtenir une meilleure approximation, à condition que $\hat{\underset{\sim}{\Gamma}}$ soit complètement connu, en considérant $X_p^2(S) = X_p^2(\hat{\delta}_.)/(1 + \hat{C}^2)$ comme ayant une distribution de χ_ν^2, où $\nu = IJ/(1 + \hat{C}^2)$, $C^2 = \sum_{k=1}^{IJ} \hat{\delta}_k^2/(\hat{\delta}_.^2\ IJ) - 1$. On peut exprimer $\hat{C}^2$ comme une fonction des estimations des proportions et des covariances au moyen du résultat

$\sum_{k=1}^{IJ} \hat{\delta}_k^2 = \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} \sum_{m=1}^{I+1} \sum_{\ell=1}^{J+1} \hat{\sigma}_{ij,m\ell}(\hat{h})/(\hat{p}_{i+}\ \hat{p}_{+j}\ \hat{p}_{m+}\ \hat{p}_{+\ell})$ sans évaluer chacun des $\hat{\delta}_i$, où $\hat{\sigma}_{ij,m\ell(\hat{h})}$ est l'estimation de la covariance entre $\hat{h}_{ij}$ et $\hat{h}_{m\ell}$.

La variable de Wald tient compte du plan d'échantillonnage et s'écrit

$$X_W^2 = \hat{\underset{\sim}{h}}^T \hat{\underset{\sim}{\Gamma}}^{-1} \hat{\underset{\sim}{h}}$$

où $\hat{\underset{\sim}{\Gamma}}$ est l'estimation de la matrice des covariances de $\hat{\underset{\sim}{h}}$ dans le plan d'échantillonnage utilisé.

On peut calculer le seuil de signification asymptotique de X_p^2 à l'aide de l'approximation de Satterthwaite :

$$\begin{aligned} SS\ (X_P^2) &= Pr[X_P^2 \geq \chi_{IJ}^2(\alpha)] \\ &\doteq Pr[\chi_\nu^2 \geq \chi_{IJ}^2(\alpha)/\{\hat{\delta}_\cdot(1+\hat{C}^2)\}] \end{aligned}$$

où $\chi_{IJ}^2(\alpha)$ est la valeur d'une variable aléatoire khi-carré à IJ degrés de liberté au début d'une région critique qui couvre un pourcentage α de la surface à la droite de la distribution. L'approximation de Solomon-Stephens conduit à l'estimation suivante du seuil de signification :

$$\begin{aligned} SS(X_P^2) &= Pr[X_P^2 \geq \chi_{IJ}^2(\alpha)] \\ &\doteq Pr[\chi_t^2 \geq \{\chi_{IJ}^2(\alpha)/A\}^{1/r}] \end{aligned}$$

où X_p^2 est ajusté à la distribution de $A\ Y^r$, Y ayant une distribution de χ_t^2.

Certains auteurs ont proposé des variables khi-carré modifiées de type $X_p^2(\hat{b}_\cdot) = X_p^2/\hat{b}_\cdot$, où $\hat{b}$ est une combinaison linéaire appropriée pour l'estimation des effets du plan d'échantillonnage dans chaque case $\hat{d}_{ij}$. Si l'on utilise l'approximation de Satterthwaite, l'estimation du seuil de signification de $X_p^2(\hat{b}_\cdot)$ est :

$$SS[X_p^2(\hat{b}_\cdot)] = Pr[X_p^2(\hat{b}_\cdot) \geq \chi_{IJ}^2(\alpha)]$$

$$\doteq Pr[\chi_\nu^2 \geq \hat{b}_\cdot \chi_{IJ}^2(\alpha)/\{\hat{\delta}_\cdot(1+\hat{C}^2)\}] .$$

Si l'on utilise l'approximation de Solomon-Stephens, l'estimation du seuil de signification est :

$$SL[X_p^2(\hat{b}_\cdot)] = Pr[X_p^2(\hat{b}_\cdot) \geq \chi_{IJ}^2(\alpha)]$$

$$\doteq Pr[\chi_t^2 \geq \{\hat{b}_\cdot \chi_{IJ}^2(\alpha)/A\}^{1/r}] .$$

Deux cas particuliers de $X_p^2(\hat{b}_\cdot)$ sont :

$$X_p^2(\hat{d}_\cdot) = X_p^2/\hat{d}_\cdot$$

$$X_p^2(\hat{\lambda}_\cdot) = X_p^2/\hat{\lambda}_\cdot$$

où

$$\hat{d}_\cdot = \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} \hat{d}_{oij}/[(I+1)(J+1)]$$

et

$$\hat{\lambda}_{\cdot} = \sum_{i=1}^{I+1} \sum_{j=1}^{J+1} (1 - \hat{p}_{i+}\,\hat{p}_{+j})\,\hat{d}_{oij} / [(I+1)(J+1) - 1]$$

où $\hat{d}_{oij}$ est l'estimation de l'effet du plan associé à $\hat{p}_{ij}$ selon H_o; autrement dit : $\hat{d}_{oij} = \mathrm{var}(\hat{p}_{ij}) / [\hat{p}_{i+}\,\hat{p}_{+j}(1 - \hat{p}_{i+}\,\hat{p}_{+j})/n]$.

Une modification effectuée à partir de l'approximation de Satterthwaite est :

$$X_p^2(S,\alpha) = X_p^2 \left[\frac{\chi^2_{IJ}(\alpha)}{\hat{\delta}_{\cdot}(1+\hat{C}^2)\ \chi^2_{\nu}(\alpha)} \right]$$

où $SS[X_p^2(S,\alpha)] = P[X_p^2(S,\alpha) \geq \chi^2_{IJ}(\alpha)]$

$\doteq \alpha$.

L'approximation de Solomon-Stephens permet d'écrire :

$$X_p^2(ST) = (X_p^2/A)^{1/r}$$

et

$$X_p^2(ST,\alpha) = X_p^2(ST)\,[\chi^2_{IJ}(\alpha)/\chi^2_t(\alpha)]$$

de sorte que $SS[X_p^2(ST,\alpha)] \doteq \alpha$.

5. Résultats empiriques

Nous examinons ici des résultats de la variable X_P^2 et de ses formes modifiées, qui ont été obtenus à partir de grands échantillons dans des tests d'indépendance appliqués à des données de l'Enquête Santé Canada. Cette enquête est fondée sur un plan d'échantillonnage par grappes stratifié à plusieurs degrés (voir La santé des Canadiens, p. 229-233). Tous les résultats sont basés sur des proportions calculées avec des totaux rectifiés pour les groupes d'âge et de sexe; une description plus détaillée des méthodes d'estimation utilisées a été présentée par Hidiroglou et Rao (1981).

- Exemple 1 : Test d'indépendance dans un tableau 4 x 2

Prenons le cas d'estimations de l'utilisation de médicaments en fonction de la consommation (quatre catégories : 0, 1, 2, 3+ classes de médicaments sur une période de deux jours) recoupées selon le sexe (hommes, femmes). Ici, n = 31,688, I+1 = 4 et J+1 = 2 (IJ = 3). Les renseignements sur la consommation de médicaments englobent des données sur l'utilisation de remèdes, de pilules ou d'onguents pendant deux jours. Les estimations du total (en milliers) de la population dans chaque catégorie de consommation et de sexe figurent au tableau ci-dessous.

Tableau 1 : Population totale selon la variété de médicaments consommés et le sexe (en milliers)

Sexe	Consommation de médicaments			
	0	1	2	3+
Hommes	6758	3091	1100	476
	(0.011)a	(0.026)	(0.027)	(0.045)
	(1.56)b	(3.37)	(1.15)	(1.38)
Femmes	5243	3659	1669	103
	(0.020)	(0.023)	(0.034)	(0.036)
	(3.59)	(3.13)	(2.85)	(1.96)

Note : "a" représente le coefficient de variation de la case
"b" représente l'effet du plan dans la case

Tableau 2 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05

Variable	Méthode d'estimation du seuil de signification	
	Satterthwaite	Solomon-Stephens
$X^2_p = 774$	0.226	0.216
$X^2_p(\hat{\delta}_\cdot) = 437$	0.062	0.060
$X^2_p(\hat{\lambda}_\cdot) = 331$	0.022	0.024
$X^2_p(\hat{d}_\cdot) = 327$	0.023	0.023
$X^2_p(S,0.05) = 408$	0.050	0.049
$X^2_p(ST) = 189$	-	-
$X^2_p(ST,0.05) = 138$	-	0.050
$X^2_W = 538$	-	-

Note : Les tirets indiquent qu'aucun calcul n'a été effectué.

Le tableau 2 présente les estimations du seuil de signification calculées à partir de l'approximation de Satterthwaite et de celle de Solomon-Stephens. Dans cet exemple, les valeurs propres sont 0.9982, 1.3728 et 2.9386 avec une estimation du coefficient de variation égale à 0.47. Les constantes de l'approximation de Solomon-Stephens ont les valeurs suivantes : r = 1.38, A = 0.558 et t = 4.766. La variable de Pearson est assez élevée (774), tandis que la variable de Wald, dont la distribution est approximativement celle d'une variable aléatoire χ^2_3 selon H_o, a une valeur relativement élevée par rapport à celle des variables de Pearson corrigées. La modification de $\hat{\delta}_.$ produit une estimation du seuil de signification qui est très proche du seuil critique visé (0.05). Toutefois, les formes modifiées de $\hat{\lambda}_.$ et de $\hat{d}_.$ produisent des seuils assez bas. En outre, les estimations du seuil de signification de l'approximation de Solomon-Stephens ou de Satterthwaite sont à peu près égales.

Par contre, la variable khi-carré obtenue pour l'approximation de Solomon-Stephens est beaucoup plus basse (189) que les autres variables. Chose curieuse, $X^2_p(ST) = 189$ et cette valeur est donc moins élevée que X^2_p/(valeur propre maximum) = 263.4. Ce résultat est contraire à ce qu'on devrait observer puisque $X^2_p = \sum_{k=1}^{IJ} \delta_{ok} W_k \leq \delta \max(\sum_{k=1}^{IJ} Wk)$, où les Wk sont des variables aléatoires indépendantes de type $\chi^2_{(1)}$ et $\delta_{max} = \max\{\delta_{ok} : k=1, 2, \ldots, IJ\}$. Or $(X^2_p/\delta_{max}) \leq \sum_{k=1}^{IJ} W_k = \chi^2(IJ)$; donc, $P[X^2_p/(\delta_{max}) \geq \chi^2_{IJ}(\alpha)] \leq \alpha$. L'explication de ce phénomène est que

l'approximation de Solomon-Stephens n'est pas exacte aux extrémités de la distribution et que le nombre $X_p^2 = 774$ est très grand et correspond à une valeur infime de α .

Une approximation basée sur trois moments de la distribution (le premier, le deuxième et le troisième) est assez bonne dans l'ensemble, mais pas nécessairement aux extrémités. Toutefois, l'analyse de Box (1954) a démontré que l'approximation de Satterthwaite est raisonnablement exacte à l'extrémité droite. Dans notre exemple, $X_p^2(S,0.05)$ aurait été égal à $X_p^2(ST,0.05)$ si X_p^2 avait été égal à 15.1. On obtient ce résultat en supposant que $X_p^2(S,0.05)$ et $X_p^2(ST,0.05)$ sont égaux, pour ensuite calculer X_p^2 :

c'est-à-dire en calculant $X_p^2 = \{\exp \frac{r}{r-1} [\ln d - \frac{1}{r} \ln A - \ln g]\}$

où $d = \chi^2_{IJ}(\alpha)/\chi^2_t(\alpha)$ et $g = \frac{\chi^2_{IJ}(\alpha)}{\hat{\delta}_\cdot (1+\hat{c}^2)\, \chi^2_\nu(\alpha)}$ et $\alpha = 0.05$, $\nu = 2.45$, $IJ = 3$, $\hat{\delta}_\cdot = 1.77$ et $\hat{c} = 0.47$.

• Exemple 2 : (Test d'indépendance dans un tableau 7 x 2)

Prenons des estimations de la population selon la tension artérielle diastolique (sept catégories : moins de 55, 55-64, 65-74, 75-84, 85-94, 95-104 et plus de 105 mmHg) et le sexe (hommes femmes). Ici n = 5,760, I+1 = 7 et J+1 = 2. Des sujets âgés de cinq ans et plus ont subi un examen de la tension artérielle. Les estimations (en milliers) calculées à partir de cette étude sont présentées au tableau 3.

Tableau 3 : Population de 5 ans et plus selon la tension artérielle diastolique et le sexe

Sexe	Tension artérielle						
	<55	55-64	65-74	75-84	85-94	95-104	105+
Hommes	390	1307	2589	3524	1905	544	184
	(0.124)a	(0.072)	(0.065)	(0.046)	(0.070)	(0.154)	(0.203)
	(1.69)b	(1.98)	(3.35)	(2.31)	(2.87)	(3.67)	(2.11)
Femmes	484	1900	3490	2981	1270	477	44*
	(0.134)	(0.060)	(0.039)	(0.063)	(0.080)	(0.173)	(0.231)
	(2.56)	(1.94)	(1.76)	(3.70)	(2.33)	(4.03)	(0.64)

Note : "a" représente le coefficient de variation de la case
"b" représente l'effet du plan dans la case

Tableau 4 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05

Variable	Méthode d'estimation du seuil de signification	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 141$	0.307	0.290
$X_p^2(\hat{\delta}_\cdot) = 81$	0.081	0.077
$X_p^2(\hat{\lambda}_\cdot) = 56$	0.019	0.021
$X_p^2(\hat{d}_\cdot) = 56$	0.018	0.021
$X_p^2(S,0.05) = 70$	0.050	0.049
$X_p^2(ST) = 59$	-	-
$X_p^2(ST,0.05) = 36$	-	0.050
$X_W^2 = 107$	-	-

Note : Les tirets indiquent qu'aucun calcul n'a été effectué.

Le tableau 4 présente les estimations du seuil de signification calculées à partir de l'approximation de Satterthwaite et de celle de Solomon-Stephens pour la variable habituelle de Pearson X_p^2 et ses formes modifiées. On décrit brièvement quelques conclusions au sujet du tableau ci-dessus à la fin de la section des exemples.

• Exemple 3 : (Test d'indépendance dans un tableau 3 x 3)

La condition physique des sujets a été évaluée d'après leur pouls à la fin d'un exercice qui consistait à monter et à descendre deux marches pendant quelques minutes. Les sujets ont été classés en trois catégories : "niveau recommandé" (nombre de pulsations inférieur à la norme au bout de six minutes), "minimum acceptable" (fréquence du pouls inférieure à la norme au bout de trois minutes, mais supérieure à la norme fixée pour six minutes) et "inacceptable" (fréquence supérieure à la norme au bout de trois minutes d'exercice) et trois genres de fumeurs ont été définis : les fumeurs habituels, les fumeurs occasionnels et les non-fumeurs. Ici $n = 1,731$, $I+1 = 3$ et $J+1 = 3$. Les estimations (en milliers) calculées à partir des données recueillies figurent au tableau 5.

Tableau 5 : Population de 15 à 64 ans selon le niveau de condition physique et le genre de fumeur de cigarettes

Genre de fumeurs	Niveau de condition physique		
	recommandé	minimum	inacceptable
Habituels	2083 (0.083)a (4.28)b	1415 (0.091) (3.26)	138 * (0.230) (2.76)
Occasionnels	218* (0.246) (4.06)	91* (0.182) (0.86)	8 * (0.801) (1.68)
Non-fumeurs	1924 (0.081) (1.86)	940 (0.071) (1.26)	32 * (0.314) (0.77)

Note : "a" représente le coefficient de variation de la case
"b" représente l'effet du plan dans la case

Tableau 6 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05

Variable	Méthode d'estimation du seuil de signification	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 24$	0.138	0.130
$X_p^2(\hat{\delta}_.) = 18$	0.071	0.068
$X_p^2(\hat{\lambda}_.) = 11$	0.008	0.009
$X_p^2(\hat{d}_.) = 10$	0.006	0.007
$X_p^2(S,0.05) = 16$	0.050	0.049
$X_p^2(ST) = 20$	-	-
$X_p^2(ST,0.05) = 14$	-	0.050
$X_W^2 = 24$	-	-

Note : Les tirets indiquent qu'aucun calcul n'a été effectué.

• **Exemple 4 : (Test d'indépendance dans un tableau 5 x 3)**

Le tableau présenté dans cet exemple recoupe des estimations de la population selon la tension artérielle diastolique (cinq catégories : 55-64, 65-74, 75-84, 85-94, 95-104 mmHg) et le genre de fumeur (trois catégories : les fumeurs habituels, les anciens fumeurs et les personnes qui n'ont jamais fumé). Ici n = 4,007, I+1 = 5 et J+1 = 3. Les estimations (en milliers) calculées pour cette ventilation figurent au tableau 7.

Tableau 7 : Population de 15 ans et plus selon la tension artérielle diastolique et le genre de fumeur de cigarettes

Genre de fumeurs	Tension artérielle				
	55-64	65-74	75-84	85-94	95-104
Habituels	824	1876	2087	943	364
	(0.096)a	(0.055)	(0.054)	(0.091)	(0.195)
	(2.19)b	(1.92)	(2.03)	(2.38)	(3.72)
Occasionnels	384	1073	1574	834	256
	(0.100)	(0.108)	(0.116)	(0.098)	(0.177)
	(1.12)	(3.99)	(5.53)	(1.98)	(2.07)
Non-fumeurs	475	1417	1817	885	293
	(0.136)	(0.066)	(0.072)	(0.064)	(0.161)
	(2.41)	(1.78)	(2.36)	(1.11)	(2.10)

Note : "a" représente le coefficient de variation de la case
"b" représente l'effet du plan dans la case

Tableau 8 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05

Variable	Méthode d'estimation du seuil de signification	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 31$	0.308	0.289
$X_p^2(\hat{\delta}_\cdot) = 19$	0.107	0.099
$X_p^2(\hat{\lambda}_\cdot) = 13$	0.023	0.024
$X_p^2(\hat{d}_\cdot) = 13$	0.022	0.024
$X_p^2(S,0.05) = 15$	0.050	0.049
$X_p^2(ST) = 24$	-	-
$X_p^2(ST,0.05) = 15$	-	0.050
$X_W^2 = 36$	-	

Note : Les tirets indiquent qu'aucun calcul n'a été effectué.

• Exemple 5 : (Test d'indépendance dans un tableau 3 x 3)

Le dernier exemple porte sur un recoupement d'estimations de la population selon le genre de boisson alcoolisée consommée (bière, vin, spiritueux) et le nombre de consommations par semaine (1-5, 6-13, 14+). Ici n = 6,966, I+1 = 3 et J+1=3. Les estimations (en milliers) figurent au tableau 9.

Tableau 9 : Population de 15 ans et plus selon le genre de boisson consommée et le nombre de consommations par semaine

Nombre de consommations	Genre de boisson		
	bière	vin	spiritueux
1-5	223 (0.082)a (1.28)b	1124 (0.038) (1.15)	4585 (0.026) (0.62)
6-13	327 (0.060) (1.00)	806 (0.049) (1.04)	777 (0.048) (1.23)
14+	13* (0.258) (0.61)	1162 (0.045) (1.94)	294 (0.063) (1.01)

Note : "a" représente le coefficient de variation de la case
"b" représente l'effet du plan dans la case

Tableau 10 : Variables khi-carré et estimations du seuil de signification, probabilité critique de 0.05

Variable	Méthode d'estimation du seuil de signification	
	Satterthwaite	Solomon-Stephens
$X_p^2 = 1974$	0.255	0.258
$X_p^2(\hat{\delta}_\cdot) = 1102$	0.072	0.072
$X_p^2(\hat{\lambda}_\cdot) = 1763$	0.210	0.210
$X_p^2(\hat{d}_\cdot) = 1765$	0.218	0.218
$X_p^2(S,0.05) = 970$	0.050	0.049
$X_p^2(ST) = 336$	-	-
$X_p^2(ST,0.05) = 282$	-	0.050
$X_W^2 = 660$	-	-

Note : Les tirets indiquent qu'aucun calcul n'a été effectué.

6. Résumé

Les valeurs de la variable khi-carré X^2_p et de quelques-unes des formes modifiées de cette variable sont résumées au tableau 11 pour les cinq ensembles de données. Ce tableau permet de noter plusieurs points. Dans tous les exemples choisis, toutes les variables khi-carré sont significatives au moins au seuil de 5 %. La variable de Wald semble être systématiquement plus élevée que les formes modifiées de la variable khi-carré de Pearson et inférieure à la variable de Pearson elle-même (voir les exemples 1 à 4).

Tableau 11 : Résumé des variables khi-carré des exemples 1 à 5

Exemple	Variable khi-carré				
	X^2_p	$X^2_p(\hat{\delta}_\cdot)$	$X^2_p(S,0.05)$	$X^2_p(ST,0.05)$	X^2_w
1. MÉDICAMENTS-SEXE	774***	437***	408***	138***	538***
2. TENSION-SEXE	141***	81***	70***	36***	107***
3. CONDITION-GENRE DE FUMEUR	24***	18***	16***	14***	24***
4. TENSION-GENRE DE FUMEUR	31***	19***	15***	15***	36***
5. GENRE DE BOISSON-NOMBRE DE CONSOM-MATIONS	1074***	1102***	970***	282***	660***

** - seuil de signification de 0.05
*** - seuil de signification de 0.01

La variable khi-carré modifiée de Solomon-Stephens, $X^2_p(ST,0.05)$ est systématiquement inférieure aux autres variables khi-carré. On note même que, quand la valeur de la variable khi-carré de Pearson est très élevée (voir les exemples 1, 2 ou 5), la variable khi-carré modifiée de Solomon-Stephens est beaucoup plus faible que les autres variables. Quand

la variable khi-carré de Pearson n'est pas élevée (comme à l'exemple 3 ou 4), la variable khi-carré modifiée de Solomon-Stephens se rapproche beaucoup de la variable modifiée de Satterthwaite, $X^2_p(S,0.05)$. Soulignons aussi que les estimations de la probabilité critique calculées à partir des approximations de Satterthwaite ou de Solomon-Stephens sont très proches l'une de l'autre.

Pour les variables $X^2_p(S,0.05)$, $X^2_p(ST,0.05)$ ou X^2_w, il faut calculer la matrice des covariances des estimations des proportions dans chaque cellule. Dans les petits tableaux de contingence, cela ne pose pas vraiment de difficultés, mais ce calcul devient une tâche énorme dans les grands tableaux, où le nombre de variances et de covariances augmente en fonction du carré du nombre de cases, qui lui-même augmente en fonction du produit du nombre de catégories prévue pour chaque variable de classification. En outre, dans le cas de X^2_w, il faut invertir la matrice des estimations des covariances, matrice qui peut être instable, tandis que pour la variable $X^2_p(ST,0.05)$ on doit calculer les valeurs propres de la matrice des effets du plan d'échantillonnage $\underset{\sim}{D}_o$. Signalons que la valeur et le seuil critique de $X^2_p(\hat{\delta}_.)$ sont proches de ceux de la variable $X^2_p(S,0.05)$ et que $X^2_p(\hat{\delta}_.)$ exige beaucoup moins de calculs.

Dans tous les exemples présentés dans la présente étude, les inférences fondées sur toutes les méthodes utilisées conduisent au même résultat : l'hypothèse nulle est rejetée. L'acceptation ou le rejet de l'hypothèse nulle à partir de la variable $X^2_p(S,0.05)$ dépend de deux facteurs, $\hat{\delta}_.$ et X^2_p. Il existe probablement des cas où les variables modifiées ne produisent pas les mêmes inférences (dans des cas limites). Il serait

intéressant de comparer le comportement de toutes les variables khi-carré proposées ici au moyen d'une analyse de Monte Carlo. Une telle étude fournirait des renseignements sur les seuils de signification réels par opposition aux seuils théoriques pour chaque variable dans diverses conditions. On pourrait poser des questions comme : À quel point les variables khi-carré et les inférences faites à l'aide de ces variables sont-elles sensibles à la variabilité des effets du plan d'échantillonnage dans chaque case? Quel est l'effet des valeurs élevées ou faibles de X_p^2 sur les inférences? Le pouvoir des inférences basées sur les petits échantillons serait aussi une question importante.

7. Remerciements

L'auteur tient à remercier avec gratitude Philip Giles pour sa collaboration à la production de bandes sommaires, ainsi que Nanjamma Chinnappa et Geoffrey Hole pour leur appui. L'auteur veut également exprimer sa reconnaissance au professeur J.N.K. Rao pour ses suggestions et ses remarques.

ANNEXE

Tableau 12 : Renseignements sommaires sur les exemples 1 à 5

Variable	Exemple 1 MÉDICAMENTS-SEXE	2 TENSION-SEXE	3 CONDITION-GENRE DE FUMEUR	4 TENSION-GENRE DE FUMEUR	5 GENRE DE BOISSON-NOMBRE DE CONSOMMATIONS
A	0.560	0.075	0.233	0.062	0.693
r	1.38	1.85	1.54	1.94	1.37
t	4.77	13.70	6.98	14.63	5.17
$\hat{d}_.$	2.37	2.50	2.31	2.45	1.09
$\hat{\lambda}_.$	2.34	2.49	2.18	2.43	1.12
$\hat{\delta}_.$	1.77	1.74	1.28	1.61	1.79
$cv(\hat{\lambda}_i)$	0.47	0.75	0.63	1.08	0.69
ν	2.45	3.84	2.86	3.67	2.71
IJ	3.00	6.00	4.00	8.00	4.00
n	31688	5760	1731	4007	6966

Tableau 13: Valeurs propres des examples 1 à 5

Exemple	Valeurs propres $(\hat{\delta}_i)$ 1	2	3	4	5	6	7	8
1. MÉDICAMENTS-SEXE	1.00	1.37	2.94					
2. TENSION-SEXE	0.51	0.73	0.82	1.80	2.19	4.32		
3. CONDITION-GENRE DE FUMEUR	0.45	0.75	1.36	2.56				
4. TENSION-GENRE DE FUMEUR	0.20	0.42	0.47	0.69	0.92	1.20	3.65	5.36
5. GENRE DE BOISSON-NOMBRE DE CONSOMMATIONS	0.26	1.15	2.17	3.58				

Tableau 14: Distribution de fréquence des coefficients de variation dans chaque, case, exemples 1 à 5

Exemple	Intervalles du coefficient de variation				Moyenne (cv)	Taille de l'échantillon
	[0-5]	[5-10]	[10-20]	(20+		
1. MEDICAMENTS-SEXE	100%	-	-	-	3%	31,688
2. TENSION-SEXE	14%	43%	29%	14%	11%	5,760
3. CONDITION-GENRE DE FUMEUR	-	57%	11%	44%	23%	1,731
4. TENSION-GENRE DE FUMEUR	-	60%	40%	-	10%	4,007
5. GENRE DE BOISSON-NOMBRE DE CONSOM-MATIONS	56%	33%	-	11%	7%	6,966

Bibliographie

BOX, G.E.P. (1954), "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification", Ann. Math. Stat. 25, 290-302.

FELLEGI, I.P., "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples", Journal of the American Statistical Association 75, 261-268.

HIDIROGLOU, M.A. et RAO, J.N.K. (1981), "Tests khi-carré pour l'analyse d'observations classées de l'Enquête Santé Canada", document présenté à la 43ieme conférence de l'IIS à Buenos Aires.

HOLT, D., SCOTT, A.J. et EWINGS, P.O. (1980), "Chi-Squared Tests with Survey Data", Journal of the Royal Statistical Society, Sec. A, 143, 302-320.

IMHOF, J.P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables", Biometrika, 417-426.

JENSEN, Donald R. et SOLOMON, Herbert (1972), "A Gaussian Approximation to the Distribution of a Definite Quadratic Form", Journal of the American Statistical Association 67, 898-902.

La Santé des Canadiens (1981), Rapport de l'Enquête Santé Canada, publication du ministère de la Santé et du Bien-être social et de Statistique Canada (nº 82-538F au catalogue).

RAO, J.N.K. et SCOTT, A.J. (1981), "The Analysis of Categorical Data from Complex Surveys : Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables", Journal of the American Statistical Association 76, 221-230.

SATTERTHWAITE, F.E. (1946), "An Approximate Distribution of Estimates of Variance Components", Biometrics 2, 110-114.

SOLOMON, H. et STEPHENS, M.A. (1977), "Distribution of a Sum of Weighted Chi-Squared Variables", Journal of the American Statistical Association 72, 881-885.